Report: Thermodynamic Neural Network (RNAP Only)

(Dated: June 24, 2019)

I. INTRODUCTION

The purpose of this report is to describe a neural network model that is equivalent to a thermodynamic model of transcription factors binding to DNA. We would like to infer the parameters of a thermodynamic model in which there are two states: (i) no transcription factor is bound to sequence so there is no transcription rate (ii) RNA polymerase is bound to sequence which leads to a transcription rate τ. The microstates of thermodynamic model we are considering are shown in Fig. 1.

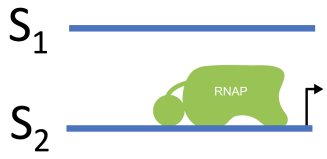


FIG. 1: Microstates of the thermodynamic model. Blue line indicates sequence, S_i represents microstate i.

The partition function for this system is

$$Z = 1 + e^{-(\epsilon_r - \mu)}, \tag{I.1}$$

where ϵ_r is the binding energy of RNAP to DNA, and μ is the chemical potential (I have set $RT = 1$).
The probabilities of the 2 microstates are

$$S_1 = \frac{1}{Z} \tag{I.2}$$

$$S_2 = \frac{e^{-(\epsilon_r - \mu)}}{Z} \tag{I.3}$$

$$\tag{I.4}$$

Assuming that the rate of transcription τ (at the sequence we are considering i.e. *lac* promoter) is proportional to the occupancy of the RNAP at its binding in thermal equilibrium, we can write down a model for the transcription rate:

$$\tau = \tau_{\max} \frac{e^{-(\epsilon_r - \mu)}}{1 + e^{-(\epsilon_r - \mu)}} \tag{I.5}$$

The goal is to train a neural network on mutagenized sequences (input) and their transcription rates (outputs) to infer the parameters θ_r (the PSAM), $\tau_{\max}$, and μ. I set parameters $\tau_{\max} = 1$ and $\mu = 2.7$ The PSAM for RNAP to generate input data was taken from Forcier et al 2018. The architecture of the neural network to do this task is shown in Fig. 2.

II. RESULTS

A histogram of the input labels (τ) can be seen in Fig. 3
The loss function used is mean squared error. Log loss vs. epochs are shown in Fig. 4a. Predictions vs. labels on the test set are shown in Fig. 4b. The PSAM used to generate ϵ_r and the inferred PSAM can be seen in Fig. 5.
The value of $\tau_{\max}$ inferred from the weight in the last layer was 1.0000634. I infer μ as follows: I take the difference between the value of the ϵ_r node and the energy values used to generate the labels (τ); I obtain 2.6999.

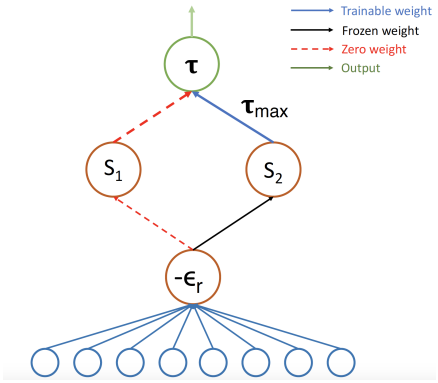
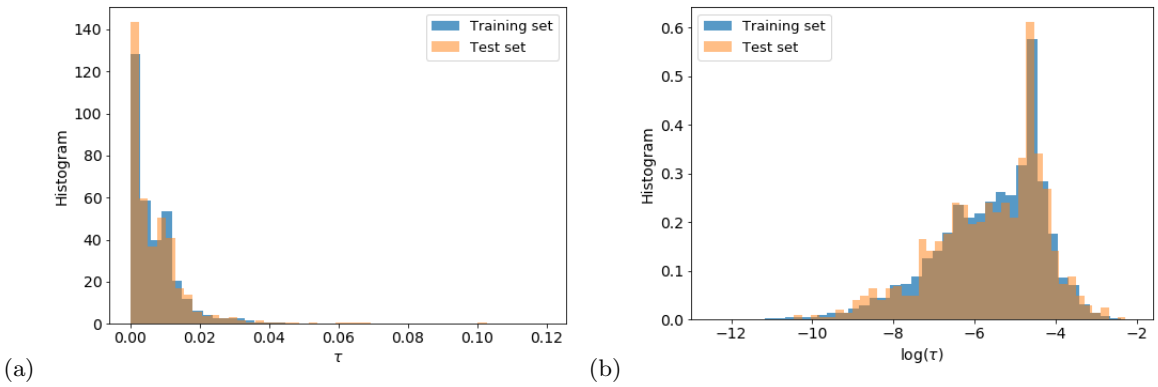
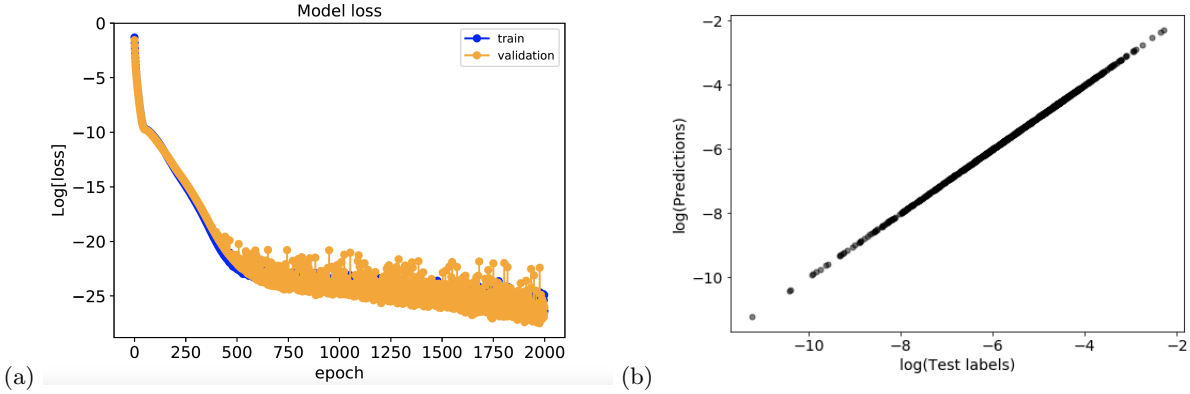


FIG. 2: Architecture of the neural network used to infer model parameters θ_r, $\tau_{\max}$, and μ.



(a) (b)

FIG. 3: Histograms of (a) τ and (b) $\log\tau$. τ was the output labels used in training the network.



(a) (b)

FIG. 4: (a) Log loss versus epochs. (b) log Predictions vs. log τ.

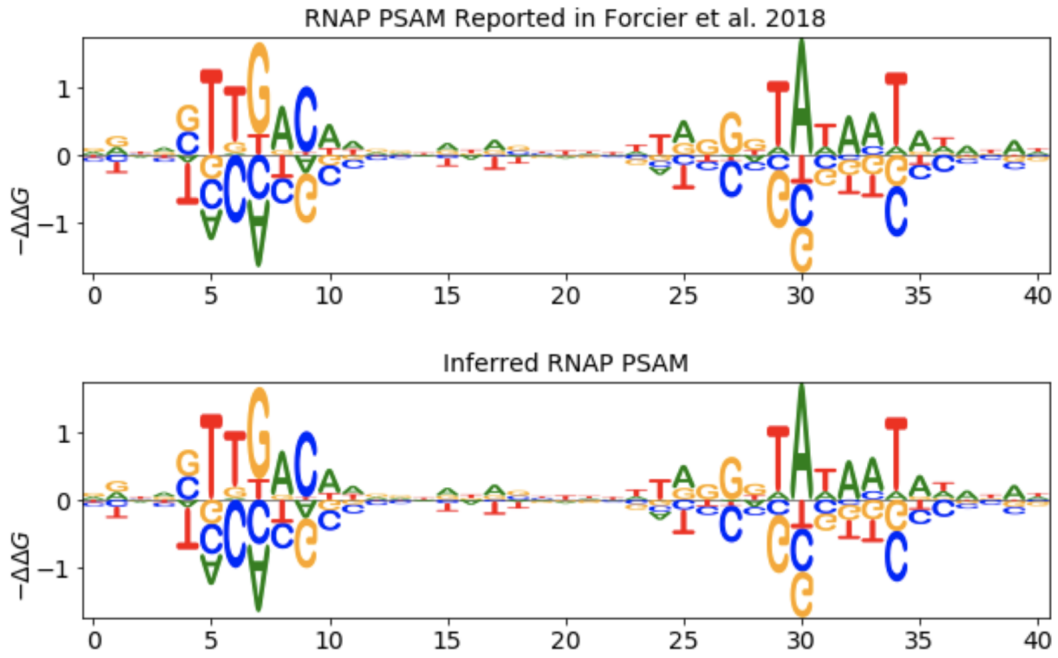


FIG. 5: (top): PSAM taken from Forcier et al 2018 (bottom): PSAM inferred from neural network weights.